

Lékař statistikem, nebo spíše statistika pro lékaře?

RNDr. Arnošt Komárek, Ph.D., Mgr. Marian Rybář

Katedra pravděpodobnosti a matematické statistiky MFF UK, Praha

Článek poskytuje lékařům cenné rady při statistickém zpracování dat v rámci jejich klinického výzkumu. Shrnuje nejčastější statistické chyby, kterých se může lékař dopustit při srovnání dvou léčebných postupů. Navrhuje také, co je možné zlepšit, pokud by se již samotná data od počátku klinického výzkumu sbírala správným způsobem.

Klíčová slova: statistika, zpracování dat, statistické metody, lékařský výzkum, klinické studie.

Doctor as a statistician or better statistics for doctors?

This article gives doctors valuable pieces of advice when processing the statistical data within their clinical research. It summarizes the most common statistical mistakes which the doctors are likely to make when comparing two methods of treatment. Moreover, the article also suggests ways of improvement if the data were collected in the appropriate way from the beginning of the clinical research.

Key words: statistica, data processing, statistical methods, clinical research, clinical studies.

Interní Med. 2010; 12(2): 114–116

Nejeden lékař se během své praxe setká se zajímavými a často i pracně sebranými daty, která je nutno vhodným způsobem zpracovat. Častým přáním je následná publikace výsledků založených na těchto datech v kvalitním odborném časopise. S rychlým rozvojem počítačů a související dostupností výpočtů nestatistikům začaly růst nároky odborných medicínských časopisů na kvalitu statistického zpracování prezentovaných výsledků. Zdánlivá jednoduchost použití moderních statistických programů přitom svádí ke sběru méně smysluplných dat, nebo dokonce chybnému použití jednotlivých statistických metod. V důsledku toho mohou být vyvozené závěry naprosto zavádějící.

V našem článku uvážíme v medicíně častý problém, a to srovnání dvou léčebných postupů, u nichž je efekt léčby měřen hodnotou kvantitativního ukazatele Y (např. krevní tlak apod.). Ukážeme, že i k takto základnímu problému lze přistupovat několika způsoby, přičemž použité statistické metody budou zahrnovat pouze jednovýběrový, dvouvýběrový a párový t-test, jež jsou známy i příležitostnému uživateli statistiky. Výpočetní stránku těchto statistických postupů přenecháme učebnicím statistiky, respektive budeme předpokládat, že techniku výpočtu zajišťuje vhodný statistický software, např. R, SAS, S-Plus, SPSS, STATISTICA.

Zaměříme se na problémy, jejichž správné vyřešení nám žádné tlačítko ve statistickém programu nezajistí. Upozorníme na chyby, které se mohou u zmíněných základních metod vyskytnout, případně co je možné zlepšit, pokud by se již samotná data sbírala jiným

způsobem. Pro potřeby článku předpokládáme pouze znalost základních pojmů (průměr, směrodatná odchylka, rozptyl, korelace, medián, nulová a alternativní hypotéza, hladina významnosti statistického testu, P-hodnota, normální rozdělení), a měl by tedy být srozumitelný i čtenáři, který se se statistikou setkává pouze při četbě výsledků studií publikovaných v odborných časopisech.

Následující 3 oddíly porovnávají tři možné přístupy ke srovnání dvou léčebných postupů:

1. jednovýběrový problém
2. dvouvýběrový problém
3. párový problém

Srovnání průměru kvantitativního ukazatele s obecně uznávanou hodnotou (jednovýběrový problém)

Představme si následující situaci: Zavedený lék či postup vede dle odborné literatury po roční aplikaci ve skupině pacientů k poklesu kvantitativního ukazatele Y v průměru o 2 jednotky. Zkoušíme nový léčebný postup a chceme prokázat, že je lepší než léčebný postup zavedený. Na základě měření u našich pacientů chceme prokázat, že populační průměr poklesu ukazatele Y je vyšší než 2 jednotky. Populačním průměrem (střední hodnotou) se rozumí hypotetický průměr z populace všech potenciálních pacientů.

Testovat budeme následující nulovou (H_0) a alternativní hypotézu (H_1):

H_0 : střední hodnota poklesu ukazatele $Y = 2$

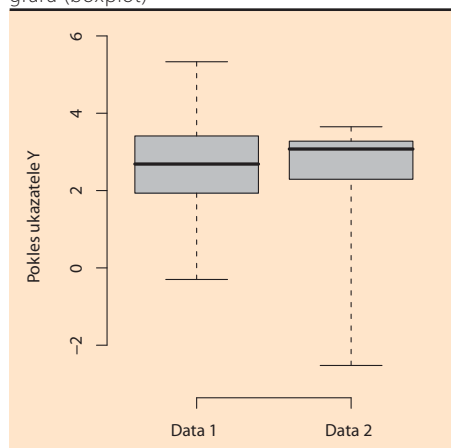
H_1 : střední hodnota poklesu ukazatele $Y > 2$

Řekněme, že v testovací skupině 15 pacientů vedla aplikace nového postupu k průměrnému poklesu o 2,6 jednotky se směrodatnou odchylkou poklesu 1,2 jednotky. Výše uvedené hypotézy se klasicky testují jednovýběrovým t-testem, jenž zde vede k P-hodnotě 0,037. Porovnání této P-hodnoty s hodnotou 0,05 (v medicíně nejčastěji používaná 5% hladina významnosti) vede k zamítnutí nulové hypotézy, a můžeme tedy konstatovat, že nový léčebný postup je statisticky významně lepší než postup zavedený.

Problém 1: t-test použitý na data s nenormálním rozdělením

Základním předpokladem pro použití jednovýběrového t-testu při malém rozsahu výběru (zde uvažujeme pouze 15 pacientů) je normální (Gaussovo) zvonovité rozdělení sledovaného znaku, v našem případě tedy poklesu ukazatele Y . K jednoduchému grafickému znázornění a hrubému posouzení normality dat lze použít krabíčkového grafu – boxplot. U normálního rozdělení je přibližně symetrický kolem středu krabíčky – obvykle mediánu. Obrázek 1 ukazuje krabíčkové grafy pro dvě sady hypotetických dat, u kterých je v obou případech průměr roven 2,6 a směrodatná odchylka 1,2. V případě první sady dat se použití t-testu nezdá nesprávným. V případě druhé sady dat je při uvažovaném rozsahu výběru 15 pacientů použití t-testu hrubou chybou, jejímž důsledkem může být chybné prohlášení nového léčebného postupu za lepší, než je léčebný postup zavedený. V praxi bohužel není výjimkou t-test provedený s daty podobného

Obrázek 1. Ukázka normálně (Data 1) a nenormálně (Data 2) rozdělených dat dle krabičkového grafu (boxplot)



rozsahu, aniž by bylo zjišťováno, zda lze normální rozdělení předpokládat či nikoliv.

Problém 2: Zbytečné odmítnutí t-testu

Na druhou stranu, mnohý uživatel statistiky se zcela zbytečně zabývá ověřováním nepodstatných předpokladů použité metody. Vrátili-li se k výše uvedenému t-testu, tak v případě, že by pacientů zahrnutých do studie bylo mnohem více než 15, nemuselo by použití t-testu být chybné ani v případě druhé sady dat z obrázku 1, kdy rozdělení zjevně normální není (statistik by měl být schopen posoudit, kolik pozorování je potřeba, aby na taková data šel použít t-test). Při zaslání požadavku na provedení t-testu dále leckterý statistický software automaticky vyprodukuje též některý z tzv. testů normality. Aby komplikací nebylo málo, pro data s vyšším počtem pozorování budou tyto testy indikovat porušenou normalitu i v případě téměř ideálně normálně rozdělených hodnot s krabičkovým grafem podobným tomu pro první sadu dat na obrázku 1. V takovém případě prakticky neexistuje optimálnější postup než t-test. Pokud bychom použili „opatrnější“ neparametrickou obdobu t-testu, která nevyžaduje normalitu dat, nemusí se prokázat zlepšení, přestože má reálné opodstatnění.

Problém 3: Snaha prokazovat platnost nulové hypotézy

Změňme nyní nepatrně zadání problému. Zavedený postup vede k poklesu kvantitativního ukazatele Y v průměru o 2 jednotky, avšak jeho aplikace je příliš náročná, ať již z praktického hlediska, nebo finančně. Zkoušíme nový léčebný postup, jehož aplikace je méně náročnější, levnější apod. K tomu, aby bylo možné plošně zavést nový léčebný postup, stačí nám ukázat, že

co do léčebného efektu je shodný s postupem zavedeným, případně, že není horší než postup zavedený. Řekněme, že se pro jednoduchost budeme snažit prokázat, že nový postup vede k poklesu hodnoty ukazatele Y též v průměru o 2 jednotky. Většina uživatelů statistiky bude patrně uvažovat test následujících hypotéz:

H_0 : střední hodnota poklesu ukazatele Y = 2

H_1 : střední hodnota poklesu ukazatele Y $\neq$ 2

Při 15 pacientech zahrnutých do studie, u nichž v průměru dojde k poklesu ukazatele Y o 2,6 jednotky při směrodatné odchylce rovné 1,2 jednotky, povede t-test těchto hypotéz k P-hodnotě 0,073. Na 5% hladině významnosti tedy tentokrát nulovou hypotézu nezamítneme. Častým a bohužel chybným závěrem potom je, že nový postup se prokazatelně neliší od postupu zavedeného. Nikdy bychom však neměli dojít k závěru, že se prokázala platnost nulové hypotézy. Závěr z tohoto t-testu by tedy měl být, že nelze prokázat odlišnost nového a zavedeného léčebného postupu. Výsledek nám však neříká nic o tom, že by oba postupy byly co do terapeutického efektu statisticky prokazatelně shodné. K prokázání shody terapeutických efektů je nutné uvažovat tzv. test ekvivalence, který však překračuje jak rámec tohoto článku, tak obsah základních kurzů statistiky.

Srovnání průměrů kvantitativního ukazatele ve dvou nezávislých skupinách (dvouvýběrový problém)

V případě, že není problémem (například kvůli finančním omezením či nedostupnosti přístrojového vybavení pro zavedený postup) aplikovat v jedné skupině pacientů zavedený léčebný postup a v jiné skupině pacientů postup nový, je výhodnější koncipovat srovnání obou léčebných postupů jako dvouvýběrový problém. Řekněme, že chceme prokázat, že nový postup je lepší než postup zavedený, tj. že vede k většímu poklesu ukazatele Y než postup zavedený. Budeme předpokládat, že ve skupině A byl aplikován postup nový a ve skupině B postup zavedený. K prokázání, že nový postup je lepší, můžeme srovnat průměry v obou skupinách pomocí dvouvýběrového t-testu s následujícími hypotézami:

H_0 : střední hodnota poklesu ukazatele Y ve skupině A = střední hodnota poklesu ukazatele Y ve skupině B

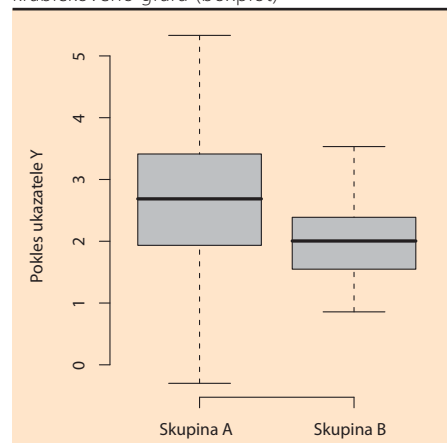
H_1 : střední hodnota poklesu ukazatele Y ve skupině A > střední hodnota poklesu ukazatele Y ve skupině B

Výhodou dvouvýběrového přístupu oproti přístupu jednovýběrovému je možnost zajistit ve skupinách pacientů (A a B) homogenitu vzhledem k dalším faktorům (věk, stupeň závažnosti choroby apod.). Tyto faktory by mohly kromě léčebného postupu ovlivnit výši terapeutického efektu. Řekněme, že v naší studii jsme nový postup aplikovali na 15 pacientech skupiny A a na 15 pacientech skupiny B. Ve skupině A jsme zjistili průměrný pokles ukazatele Y o 2,6 jednotky se směrodatnou odchylkou poklesu rovnou 1,2 jednotky, ve skupině B jsme pozorovali průměrný pokles ukazatele Y o 2 jednotky se směrodatnou odchylkou poklesu 0,64 jednotky. Dvouvýběrový t-test poté vede k P-hodnotě 0,049. Na 5% hladině významnosti se nám tedy těsně podařilo prokázat, že nový postup je lepší než ten zavedený.

Problém 4: Dvouvýběrový t-test použitý na data s nestejnými rozptyly v porovnávaných skupinách

Podobně jako jednovýběrový t-test, též dvouvýběrový t-test předpokládá normální rozdělení sledovaného znaku v obou porovnávaných skupinách, přičemž při dostatečně vysokém počtu pozorování v jednotlivých skupinách lze tento předpoklad do jisté míry pominout. Klasická forma dvouvýběrového t-testu probíraná v prakticky každém základním kurzu statistiky dále předpokládá, že variabilita (rozptyl) sledovaného znaku (v našem případě poklesu ukazatele Y) je shodná v obou porovnávaných skupinách. Splnění tohoto předpokladu lze graficky posoudit pohledem na krabičkové grafy (obrázek 2). V případě shody rozptylů by obě krabičky a délky „fousů“ měly být přibližně stejné. V našem případě tomu tak není, a je tedy důvodné podezření, že rozptyly shodné nejsou.

Obrázek 2. Ukázka porušení předpokladu o shodě rozptylů u pacientů skupiny A a skupiny B dle krabičkového grafu (boxplot)



Porušení tohoto předpokladu může mít opět za následek neoprávněné prohlášení nového postupu za lepší, než je postup zavedený. Nicméně již před 70 lety bylo odvozeno zobecnění dvouvýběrového t-testu (Welchův test), který shodné rozptyly nepředpokládá, a výsledek provedeného testu je korektní i v situaci, kdy rozptyly shodné nejsou. Kupříkladu pro náš ilustrativní příklad s 15 pacienty v každé skupině dosáhne Welchovo zobecnění dvouvýběrového t-testu P-hodnoty 0,051. Na 5% hladině významnosti se nám tedy na rozdíl od klasického t-testu (zde nesprávně použitého) těsně nepodařilo prokázat, že by nový postup byl lepší než postup zavedený.

Problém 5: Snaha vyvozovat závěry ze studie s nenáhodně rozdělenými skupinami

Častým problémem je snaha vyvozovat závěry o rozdílnosti zkoumaného ukazatele mezi dvěma skupinami v situaci, kdy o rozmístění pacientů do skupin rozhoduje též znalost nějakého faktoru více či méně ovlivňujícího hodnotu ukazatele našeho zájmu. Asi každému je jasné, že přiřazení pacientů s příznivou diagnózou do skupiny, u které chceme prokázat lepší léčebný efekt (nový léčebný postup), a pacientů se závažnou až beznadějnou diagnózou do skupiny, u které chceme prokázat horší léčebný efekt (zavedený léčebný postup), povede ke statistické analýze, jejíž výsledky lze označit za podvod. Klasickými faktory s významným vlivem na hodnotu sledovaného ukazatele jsou například věk, body mass index (BMI), počet cigaret vykouřených za den apod. Lékař provádějící studii, často nevědomky, motivován kupříkladu praktickými důvody (nový léčebný postup se snadněji aplikuje mladším pacientům apod.), systematicky přiřazuje mladší pacienty do jedné a starší pacienty do druhé z porovnávaných skupin. Následky takového jednání jsou prakticky stejné jako při úmyslném přiřazování pacientů s příznivou diagnózou do skupiny, jež se má díky statistické analýze ukázat jako lepší. Později přizvanému statistikovi poté nezbyvá než konstatovat, že pracně sebraná data jsou pro poctivé srovnání dvou léčebných

postupů prakticky bezcenná. Podobným problémem lze přitom relativně snadno předejít randomizací (znáhodněním), kdy o přidělení pacientů do skupin rozhoduje pouze náhoda (například losování). Náhoda zajistí, že hodnoty nejrůznějších faktorů, jež mohou mít (aniž bychom si toho byli vědomi) vliv na hodnotu ukazatele našeho hlavního zájmu, jsou v obou skupinách promíchány v přibližně stejném poměru.

Srovnání průměrů kvantitativního ukazatele ve dvou závislých skupinách (párový problém)

Párové testy jsou využívány pro srovnání dvou skupin v situaci, kdy jsme si vědomi jistých závislostí mezi oběma porovnávanými skupinami. Obecně platí, že je-li možné studii koncipovat jak párově, tak dvouvýběrově, má párová studie vyšší šanci vést k signifikantnímu výsledku než studie dvouvýběrová.

Každá dvojice měření provedených na témže pacientovi je „více závislá“ než měření provedená na různých pacientech a odsud plyne závislost skupin A a B. Vraťme se k našemu problému z předcházejícího oddílu, kdy jsme se snažili prokázat, že nový léčebný postup je lepší než starý léčebný postup, a uvažme, že ještě nemáme žádná data a jsme teprve v přípravné fázi studie. Jestliže je z praktického hlediska možné na každém pacientovi aplikovat postupně oba léčebné postupy, můžeme studii koncipovat tak, aby vedla na dvouvýběrový problém (dvě skupiny různých pacientů, u každého aplikujeme buď nový, nebo zavedený léčebný postup), nebo aby vedla na párový problém (jedna skupina pacientů, u každého pacienta nejprve aplikujeme jeden a poté druhý léčebný postup, ne nutně ve stejném pořadí u všech pacientů). Možnost aplikovat dva léčebné postupy na jednom pacientovi obvykle odpadá v případě, kdy léčebným postupem je operace. Nicméně například v situaci, kdy léčba pouze vede ke zmírnění příznaků choroby, jež se po určitém čase ve stejném rozsahu vrací, je párová studie možná. Obdobně lze párování využít v případě, kdy ošetřujeme stejný problém u jednoho pacienta (párově lze porovnávat kupříkladu dva různé typy zubních implantátů).

Závěr

Věříme, že mnohý čtenář nalezl v uvedených modelových případech paralelu s vlastními výzkumy, ať již v minulosti řešenými či teprve plánovanými. Náš článek samozřejmě postihuje pouze zanedbatelnou část možností a nástrah statistiky a je nutno podotknout, že též statistika se stejně jako jiné obory vyvíjí a nabízí stále se rozšiřující spektrum metod. Vztah statistika a lékaře je v dnešní počítačové době v mnohém podobný vztahu lékaře a pacienta, který má snadno na dosah informace libovolného druhu, nicméně bez odborné pomoci může být použití těchto informací spíše kontraproduktivní. Mnohý potencionální pacient se často samostatně léčí hledáním příznaků svého neduhu na internetu, sledováním rad tam nalezených a lékař se dostane ke slovu až v okamžiku, kdy se ze snadno léčitelného neduhu stává neduh těžko či vůbec léčitelný. Podobně jako pacient má často možnost volby mezi léčbou internem a léčbou vedenou odborníkem s rozhledem a znalostí možných souvislostí, má lékař požadující statistické vyhodnocení své studie možnost volby mezi mačkáním tlačítek statistického programu a spoluprací s profesionálním statistikem, který by měl s nadhledem umět posoudit, zda zmáčknutí konkrétního tlačítka opravdu vede k odpovědi na lékařem položenou otázku. V případě, že používaný software nenabízí vhodné tlačítko, měl by být schopen navrhnout alternativu. Čtenář nechť si nyní sám odpoví na otázku: „Lékař statistikem, nebo spíše statistika pro lékaře?“

Článek vznikl za podpory výzkumného záměru MSM 0021620839.

RNDr. Arnošt Komárek, Ph.D.

Mgr. Marian Rybář

Katedra pravděpodobnosti a matematické statistiky MFF UK,
Sokolovská 83, 186 75 Praha 8
www.karlin.mff.cuni.cz/komarek
www.grantstat.cz
arnost.komarek@mff.cuni.cz
marian.rybar@msps.cz